Methodology and Ethical Analysis of a Firearm Injury Prediction Model

Zoe Werner

2024-10-22

Many public health issues are routinely addressed by preventative healthcare, but some remain largely overlooked or undiscussed. For example, providers often screen patients for health risk factors, such as tobacco use, heavy drinking, and hypertension. These screenings allow providers to prevent or identify chronic health concerns related to these risk factors. Currently, firearm injury healthcare interventions are limited to patients that indicate thoughts of suicide. These interventions include counseling on storage, violence prevention, proper locks, and safety plans. These practices have indicated improvement in gun safety practices and reduced risk of accidental and self-influenced firearm injuries. However, from a healthcare system's standpoint, universal screening for every patient at any appointment would prove inefficient and resource-intensive. This study – *A machine-learning prediction model to identify risk of firearm injury using electronic health records data*, proposes an alternative to universal screening: a statistical model that predicts the probability of a firearm injury based on a data set of potential predictors. This model will allow providers to target high-risk patients and target firearm injury risk screenings and prevention efforts towards this group. In addition, it will increase the number of high-risk patients who receive firearm injury risk screening and improve the efficiency of such screenings. However, this study raises a few ethical concerns, such as the usage of patient data, the possibility of geographic and demographic bias, and potential exclusion of groups that don't meet the 'high-risk' classification.

The study uses information from Kaiser Permanente Southern California, or KPSC, a large healthcare system that operates in Southern California and provides medical and preventative care to over 5 million members. The service area includes urban, suburban, and rural areas, and provides care in 16 hospitals and in over 230 medical offices (Zhour, 2024). All care provided to members is recorded in the electronic health records, or EHR. The study was approved by the KPSC Institutional Review Board (IRB), which ensures its compliance with ethical standards, including protection of privacy and confidentiality. Researchers also obtained a waiver of consent, which means that securing consent was most likely not feasible and the study has minimal risk of breaching privacy. The cohort and predictor set were built using SAS Enterprise Guide 9.4 and the prediction model was built and validated using R version 4.0.4.

Researchers identified all adult patients who were members of KPSC, had at least 1 in-person healthcare encounter between 2010 and 2018, and had a documented firearm injury. Patients with no healthcare encounter three years before the index date (the date of firearm injury, or in the case of multiple, the earliest firearm injury) and patients with unknown gender information were excluded. Physician collaborators consider three years to be a clinically meaningful period of time and this time allowed a larger number of patients to be included, even if they didn't use the healthcare system more than once every three years. All patient encounter information – including primary and specialty care encounters, urgent care, emergency department visits, and hospitalizations – in the three years prior to the firearm injury was secured for qualifying patients.

The most recent 20% (990 injuries) of firearm injuries were reserved for final model testing and set aside. The remaining data was split into training data (70% – 1786 injuries) and validation data (30% – 596 injuries). Since the firearm injury prevalence is very low (0.01% of the total healthcare encounters during 2010-2018), they randomly selected 5 control patients to improve model performance with the highly imbalanced data. The control patients had not experienced a firearm injury and had a healthcare encounter within one year of each case injury date.

The predictor data set consisted of over 170 variables, including but not limited to age, gender, healthcare visits, diagnoses, neighborhood crime rates, and household income. They were compiled based on an extensive literature review, physician input, and structured or unstructured clinical provider notes. Predictors with more than 40% missing data or two predictors with a correlation of at least 0.8% were excluded from the predictor set. Information on suicidal thoughts, behaviors, and attempts recorded in unstructured clinical notes were extracted using a natural language processing (NLP) approach, and self or non-self-inflicted firearm injuries were predicted separately by conducting sensitivity analyses.

Predictor categories were sorted into predictor domains, which included current indication/history of suicidal ideation or attempts, individual-level socio-demographic characteristics, individual-level clinical and healthcare utilization variables, and census tract-level neighborhood information. Individual-level data was collected during usual care and individual-level socio-demographic and clinical data was collected based on the medical record number (MRN) unique to each patient. California state death certificate information was linked to each patient's social security number. Finally, each patient's address was geo-coded (transforming a description of a location to a location on the earth's surface) to the census tract where the patient resided.

This study used XGBoost, an "optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable (*XGBoost documentation*)." It was chosen because it has shown high performance in various situations, properly handles missing data, flexibly

accommodates non-linear associations and interactions, and does not require the analyst to specify the model structure a priori (reasoning from self-evident propositions before examination or analysis). XGBoost was used to predict the combined fatal and nonfatal injuries. To prevent overfitting and maximize efficiency, they used 5-fold cross-validation when training the model. Sensitivity, specificity, and positive predictive value (PPV) at a probability cutoff of 0.2% were used to assess the model performance.

Several hyper-parameters are used by XGBoost to accomplish best performance. 'Eta' adjusts the learning rate of the model and shrinks the weights of features after each round to prevent overfitting. 'Gamma' prevents overfitting and controls the penalty on coefficients that don't improve the performance (regularization). 'Max_depth' refers to the maximum depth of a tree and when increased, it heightens the complexity and likelihood of overfitting. Subsample, the number of samples applied to a tree, occurs once each boosting iteration and prevents overfitting. 'Colsample_by_tree' is number of variables supplied to a tree. 'Lambda' is the L2 regularization term that improves generalization and reduces overfitting. 'Scale_pos_weight' is useful for unbalanced classes and controls the balance between positive and negative weights. Finally, 'nrounds' is the number of boosting iterations, or steps required. Using Bayesian optimization, the parameters that achieved the highest accuracy were selected for the final set.

While all predictors were part of the initial prediction model, they used the gain metric – the improvement in accuracy attributed to a feature – to measure the feature importance and ranked each by contribution to prediction success. They observed variable importance drop with a gain of 0.017, which narrowed the predictor set to 15 predictors. The full set had a sensitivity of 0.84 and the 15 predictor set had a sensitivity of 0.83. Since the sensitivity difference was not drastic, researchers concluded that accuracy would not differ dramatically and chose the reduced variation as the final prediction model. After assessing overfitting possibilities and ensuring performance by applying the model to the validation data set, the test dataset was applied to the final model.

With the creation of this model, researchers aimed to identify a high-risk group and maximize the resource allocation and efficiency of firearm injury screening. The intent is to focus resources on groups that are most likely to be impacted, which will increase awareness regarding gun safety practices.The issue of data privacy was alleviated by approval of the IRB and obtaining a waiver of consent. In addition, the concentration on maximizing sensitivity while still improving accuracy decreased the chances of accidental bias based on demographic or geographic factors. However, if this performance model is enacted, groups that are considered low-risk will not receive firearm injury screening. According to the Pew Research Center, mass shootings have increased from 27 in 2010 to 61 in 2021. By focusing all resources and screenings on high-risk groups, the study neglects to acknowledge those who experience mass shootings at random. While the study is

incredibly significant when identifying high-risk homicidal or suicidal firearm injuries, it excluded low-risk groups from receiving firearm injury screenings and receiving beneficial information on proper gun safety practices.

Resources

Beginners tutorial on XGBoost and parameter tuning in R tutorials & notes: Machine learning. HackerEarth. (n.d.). https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

Gramlich, J. (2023, April 26). What the data says about gun deaths in the U.S. Pew Research Center. https://www.pewresearch.org/short-reads/2023/04/26/what-the-data-says-about-gun-deaths-in-the-u-s/

Merriam-Webster. (n.d.). A priori definition & meaning. Merriam-Webster. https://www.merriam-webster.com/dictionary/a%20priori

XGBoost documentation. XGBoost Documentation - xgboost 2.1.1 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/

XGBoost parameters. XGBoost Parameters - xgboost 2.1.1 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/parameter.html#parameters-for-tree-booster

Zhou, H., Nau, C., Xie, F., Contreras, R., Ling Grant, D., Negriff, S., Sidell, M., Koebnick, C., & Hechter, R. (2024). A machine-learning prediction model to identify risk of firearm injury using electronic health records data. Journal of the American Medical Informatics Association, 31(10), 2173–2180. https://doi.org/10.1093/jamia/ocae222